Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Forecasting Volatility in Stock Market Considering Sentiments of Macroeconomic News

Sagar Janokar

Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
sagar.janokar@vit.edu

Ritesh Pokarne

Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
ritesh.pokarne.20@vit.edu

Ankur Raut

Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
ankur.raut20@vit.edu

Tanmay Patil

Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
tanmay.patil201@vit.edu

Piyush Sonar

Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Pune, India
piyush.sonar20@vit.edu

Article Info
Page Number: 6480-6499
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract
Stock price forecasting is an important and growing topic in the financial engineering. Due to volatile and nonlinear nature of the global equity market it is very tough to forecast the accurate price of a particular stock in the equity market. And also, with the increased computational capabilities and introduction of artificial intelligence, and different programming methods the prediction of the stock price of a particular stock have proven efficient. For predicting the stock price, it requires the evaluation of huge amount of data. For that we have taken past four years stock data and the recent social media mainly the twitter data which is basically the data which is particularly affecting the price of the data in future. Final we compare our algorithmic outcome result with the actual Next's day actual close price. The penultimate goal of our project is to predict the behaviors of the stock equity market for the future via sentimental analysis of twitter dataset which contains set of different tweets over the past couple of days. The final results were promising, and we also found the correlation between the tweets, data of the particular stock and the stock price.
Keywords: Stock Market, Regression, Prediction, Machine Learning, Timeseries, Sentiments, Macroeconomic Data.

I. INTRODUCTION

One of the most trending areas of research is the stock price forecasting or prediction. The global stock market or the world's stock market contain huge wealth. It is valued around $ 89.5 trillion in 2021. For forecasting a particular stock price in the future[9], it is quite a challenging task and there are various factors on which it depends such as global economy, political condition, company's financial reports and also the performance etc[10]. Thus, every investor wants to maximize their profits and minimize losses, for that there are different techniques through which we can predict the future price of stocks by analysing their trends over the last few years, and which could be proven highly useful for making profits in the market. For this we have developed a model which based on the time-series forecasting which is used for the non-stationary data. In our model we have mainly two datasets, main datasets effecting the performance of a stock which is the social media sentimental analysis.

Nowadays, there is a huge amount of data which has information about numerous topics going in our surrounding, which is being transmitted through various social media platforms including Twitter, Instagram and many more. By an analysis there are around 400 million tweets are sent daily[11]. Here, positive refers that the impact to the stock price will be positive and due to which the stock price will increase accordingly and negative classifies the decrease in price of the stock and the neutral sentiments classifies that it neither increases nor decreases the stock price[12]. For that we have mainly used five different machine learning (ML) algorithms which are Random Forest Regression, Ridge Regression, Linear Regression, Ada-Boost Regression and the XG-boost Regression for analysing stock prediction and by analysing through different methods we also understand that which method has high accuracy and performance by comparing there RMSE (Root Mean Square Value), MAE (mean absolute error) and R2 (R squared error).

The remaining of the paper is discussed in a systematic way which is explained below. In section 2 we have described many related works of stock market forecasting using different methods of Regression. In the next section we have compared these all-related papers with our proposed system and compared them. Next Section states the methodology of our proposed approach which includes the data pre-processing of twitter and financial data, implementation flow with the model deployment, flowcharts and various data visualization graphs. In the next section we have discussed the results and conclusion of our model by comparing our different machine learning model with different parameters and analyse which algorithm is best suited for the forecasting of our stock. In the next section we have explained our proposed work with respective flowcharts. In the last section we have mentioned the scope of conclusion, propose our ideas for further future work. And also the implementation of our model.

II. RELATED WORK

"The profitability of daily stock market indices trades based on neural network predictions: Case study". In this paper they have studied and done the case study on the profitability of daily stock index trades based on the neural network predictions, for the S&P 500 , the DAX, the TOPIX and the FTSE in the period 1965 – 1999.This is a paper which describes an ANN (artificial neural network) model to forecast the daily stock market index return using the data from different stock markets. The main aim of the paper was to make profitable trading in the stock market. Their study mainly uses analysis of daily closing values of the companies like the German DAX Index, the London's financial times stock exchange index (FTSE), Standard and Poor's 500 index(S&P500) and the Japanese TOPIX index. For the performance prediction of their model they have set a benchmark of linear autoregressive model and then their neural network model is compared.[1]

"*Parameters for stock market prediction*". In this paper the authors have made a model by actually surveying nine different input parameters published in different research papers and articles. From that they have find the most significant input parameter that actually enhances the forecasting accuracy of their model. After doing the survey they found out that the Machine learning approach uses mainly technical variables instead of the fundamental variable in forecasting the stock indices, while the mostly used variables for stock market index prediction was the microeconomic variable. But, they concluded that when we use hybridized tyle variable then the produced results were better than before when only single input variable type is used.[2]

"Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". This paper analyses and compares the Indian stock equity market with four different approaches specifically SVM(support vector machine), ANN(artificial neural networks), the Random forest (RF) and the Naïve Bayes with double input approach. The first one is that they have computed the input data which is involved in stock market trading where they have mainly used ten technical parameters (open, low, high and close, price) and the second approach actually based on focusing on how they can represent these technical parameters as a change in deterministic data. Then they assessed the accuracy of their prediction model for the two different input approaches. Their model results show that , the first approach that we have used in giving input outperforms the other 3 prediction models.

When these technical parameters are represented as deterministic data this approach actually enhances the performance of their model.[3]

"Deep learning networks for stock market analysis and prediction". This paper is based on deep learning networks for the stock market prediction and analysis. This approach actually extracts features from a very big dataset which actually makes it useful for correctly predicting stock market. They have analyzed the deep learning algorithm approach with its set of advantages and drawbacks which is used for predicting stock market index. They have done analysis of three unsupervised approach for feature extraction the first one is the principal component analysis (PCA), auto encoder and Boltzmann machine on the overall network to forecast the intended behavior of the market. They have mainly used 38 companies' data from the list of the Korean KOSPI stock market from where they belong, from the period of January 4, 2010 to December 30, 2014. There results suggested that the deep neural networks approach can actually take out additional information from the linear autoregression model and also improves the predicting performance for the model.[4]

"Support Vector Machines for Prediction of Futures Prices in Indian Stock Market". Here, two machine learning (ML) models, Back Propagation Neural Network (BPN) and the Support Vector Machines (SVM) to predict future stock prices are implemented and compared. The dataset used is of the National Stock Exchange (NSE) of India Limited which contains 990 samples with 5 features between a time period of 1st January, 2007 to 31st December, 2010. The observation made is that SVM produces better results than BPN as SVM makes use of Structural Risk Minimization Principle which offers more generalization.[5]

"Stock Prediction Using Twitter Sentiment Analysis". In this paper Sentiment Analysis and ML techniques are used to find relation between public and market sentiment using the twitter sentiment dataset and predict stock prices. They have also used the Dow Jones Industrial Average (DJIA) values dataset obtained from Yahoo Finance. A cross validation technique along with Self Organizing Fuzzy Neural Networks (SOFNN) is proposed to do so. This technique is based on the one proposed by Bollen et al in his paper which proposes an accuracy of 87%.[6]

"Stock Closing Price Prediction using Machine Learning Techniques". They have described this paper a system that uses Artificial Neural Network to predict closing price on next day and Random Forest Techniques to make comparative analysis is proposed. They have taken the dataset from Yahoo Finance which consists of all the records between 4th May, 2009 to 4th May, 2019. The company whose data is used include Goldman Sachs, Johnson and Johnson, Nike, JP Morgan Chase and Co and Pfizer. Also they have added new variables to the available datasets to improve the accuracy. Results show that the model produces MAPE - 0.77, RMSE - 0.42,MBE - 0.013.[7]

"Random Forest Based Feature Selection of Macroeconomic Variables for Stock Market Prediction". This paper tries to investigate the level of importance between the different sector of stock prices and MV and predict the 30 days main stock index price using Random Forest (RF) with better leave-one-out cross-validation tactics and the Long-Short-Term Memory Recurrent Neural Network (LSTMRNN). Empirical analysis of the nominated model on the

particular Ghana Stock Exchange (GSE) shows high forecasting accuracy and better mean absolute error estimated to different time series techniques.[8]

III. Dataset Analysis

A. Macroeconomic Data

We pulled macroeconomic news from Twitter; news sources will tweet the headline and link to a companion news article that allows us to approach key topics frequently and fairly condensed format. In addition, hedge funds, banks, and analysts will often tweet either articles or brief views of the market. We selected 70 accounts we thought were relevant and tweeted with a significant frequency. The Twitter API sallows you to export the most recent 3200 tweets per account from a site that provides at least a month of tweets (and therefore messages) per account. Since we are examining the immediate market reaction and thus using intraday data, this is more than sufficient. It reached over 200,000 tweets.

i. Data Cleaning

We pulled macroeconomic news from Twitter; news sources will tweet the headline and link to a companion news article that allows us to approach key topics frequently and fairly condensed format. In addition, hedge funds, banks, and analysts will often tweet either articles or brief views of the market. We selected 70 accounts we thought were relevant and tweeted with a significant frequency. The Twitter API allows you to export the most recent 3200 tweets per account from a site that provides at least a month of tweets (and therefore messages) per account. Since we are examining the immediate market reaction and thus using intraday data, this is more than sufficient. It reached over 200,000 tweets. We consider only tweets between 9 am and 4 pm to align with the market intra-day data.

ii. Determining Subjectivity and Polarity

TextBlob is an appealing and reasonably light Python 2/3 toolbox for NLP and sentiment analysis development that allows simpler use and a less challenging learning curve. We utilised TextBlob for estimating subjectivity and polarity. Subjectivity and polarity are two characteristics of TextBlob's built-in sentiment analysis capability. The TextBlob and VADER processes are the most popular methods for analysing sentiment using TextBlob (Valence Aware Dictionary and Sentiment Reasoner). Given its form and intended use, TextBlob has few practical capabilities that set it apart from its rivals. Although it is robust and packed with features, it still depends on unimpressive external resources for performance. Additionally, we utilised polarity to determine the positive, negative, and neutral values.

iii. Exploratory Data Analysis (EDA)

Exploratory data analysis is a crucial procedure that entails completing a preliminary study of the data in order to identify patterns, spot anomalies, test hypotheses, and confirm assumptions using summary statistics and graphical representations. We have applied the same Exploratory data analysis process to the sentimental data five different stock companies (Apple, Microsoft, Tesla, Nvidia and PayPal) and observed the following incites as shown in **Error! Reference source not found.**,Figure 2 ,Figure 3, Figure 4, Figure 5.

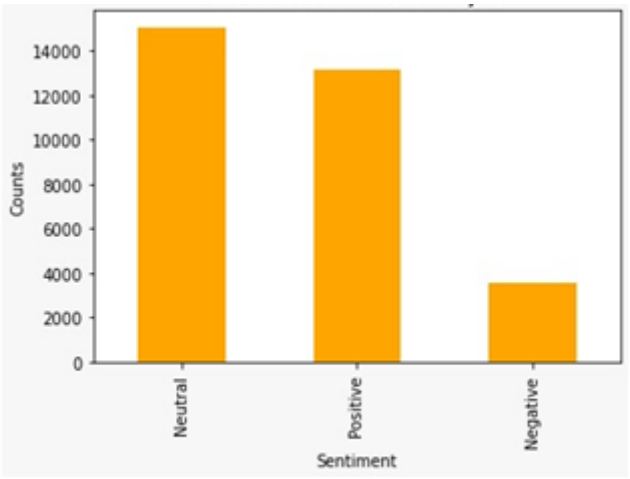

Figure 1 Apple News Sentiments Analysis

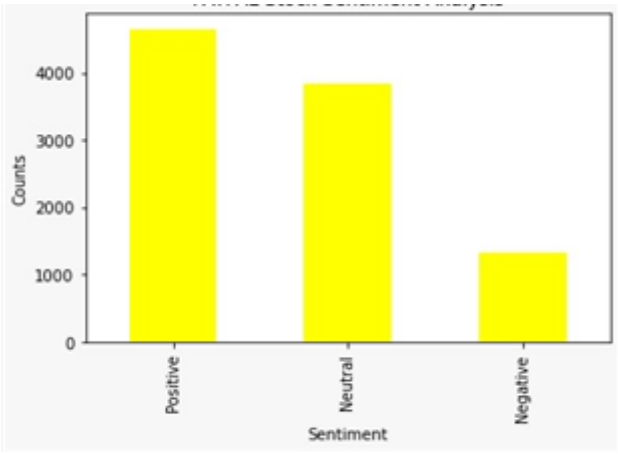

Figure 1 PayPal News Sentiments Analysis

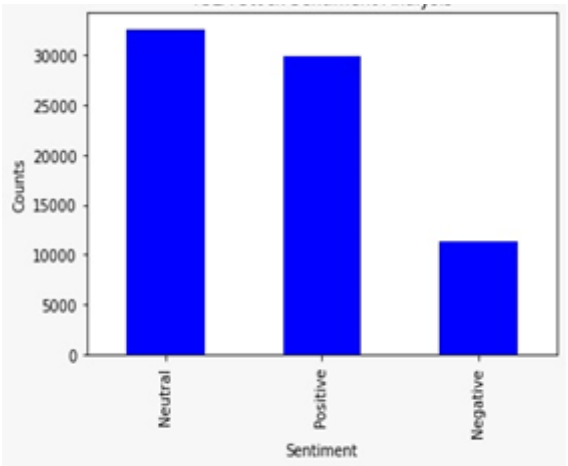

Figure 2 Tesla News Sentiments Analysis

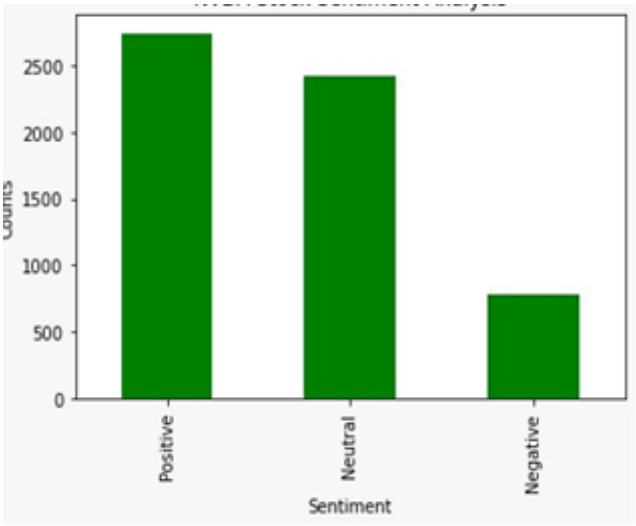

Figure 3 Nvidia News Sentiments Analysis

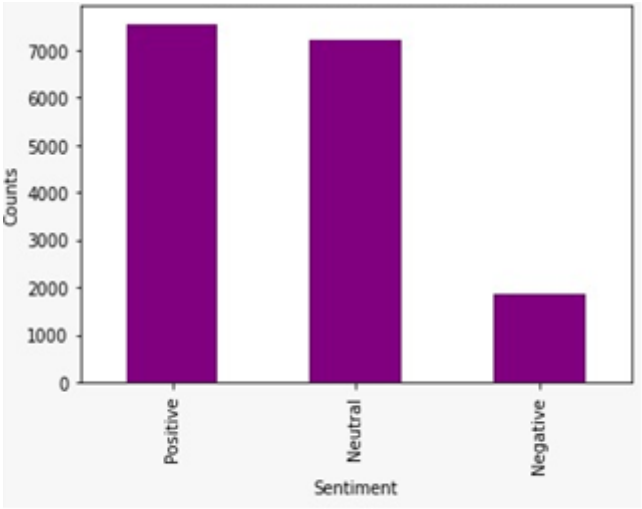

Figure 4 Microsoft News Sentiments Analysis

B. *Financial Data*

Individual data is acquired from the website finance.yahoo.com for time series study. This page displays daily statistics for the years 2000 to 2020 for the historical price development of Apple Inc., Microsoft, Tesla Inc., Nvidia, and PayPal shares. Microsoft Excel is used to process the acquired data. There are 5,283 lines in the data file. The data set must then be edited to remove unnecessary information, such as opening prices or the lowest and highest stock prices. The dataset only retains the columns with the date and closing stock prices. The examination and comparison of the collected data is the following phase.

i. Data Preprocessing

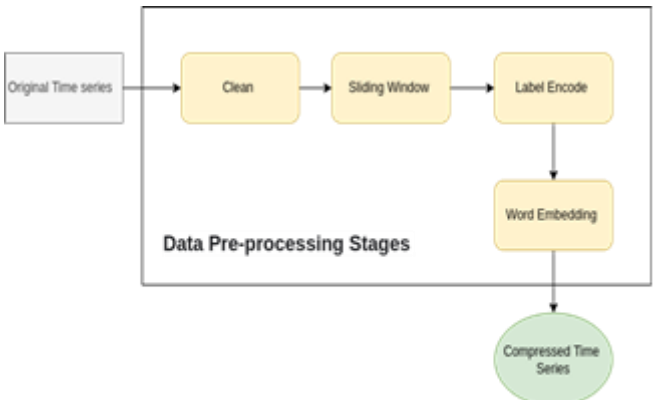

Figure 5 Data Preprocessing for Financial Data

The input time series are compressed and transformed into a form that can be used by machine learning models as part of the preprocessing phase. The pipeline phases are depicted in Figure 6. Through the use of exploratory data analysis, the precise layout of this pipeline, including the processing stages and their timing, is established (EDA). EDA focuses on comprehending correlational analysis that shows connections and other patterns as well as data exploration visualisations (such as histograms, boxplots, scatterplots, etc.). A cleanup stage in the pipeline may involve (1) finding N/A values, (2) spotting gaps, and (3) looking at outliers. Depending on the requirements of the application, affected instances can either be eliminated, replaced with interpolated values, or kept.A continuous time series is turned into discrete sequences using a sliding window method. The window's length, which determines how long the generated sequences are, and the step size can also be changed. For instance, the sequence set 1-2-3, 2-3-4, 3-4- 5, 4-5-6 will be produced by the sequence 1-2-3-4-5-6 with a sliding window size of 3 and a step size of 1. By marking metadata sequences, extra, perhaps important information about the incoming data is encoded. Although metadata labels are used to stratify or balance the data set for training and testing purposes, they are not provided into the artificial neural network model as input features.

ii. Exploratory Data Analysis (EDA)

Exploratory data analysis is a crucial procedure that entails completing a preliminary study of the data in order to identify patterns, spot anomalies, test hypotheses, and confirm presumptions using summary statistics and graphical representations. Using stock data from five other stock firms (Apple, Microsoft, Tesla, Nvidia, and PayPal), we performed the same exploratory data analysis approach and noted the trends in Figure 7, Figure 8,Figure 9, Figure 10, Figure 11.

Figure 6 Apple Stock Price Analysis

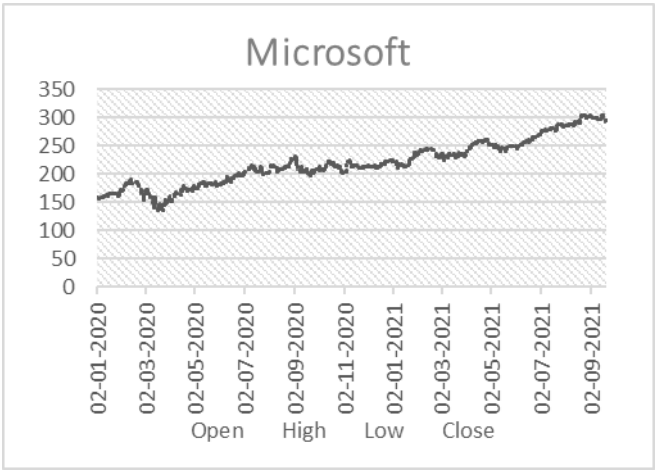

Figure 7 Microsoft Stock Price Analysis

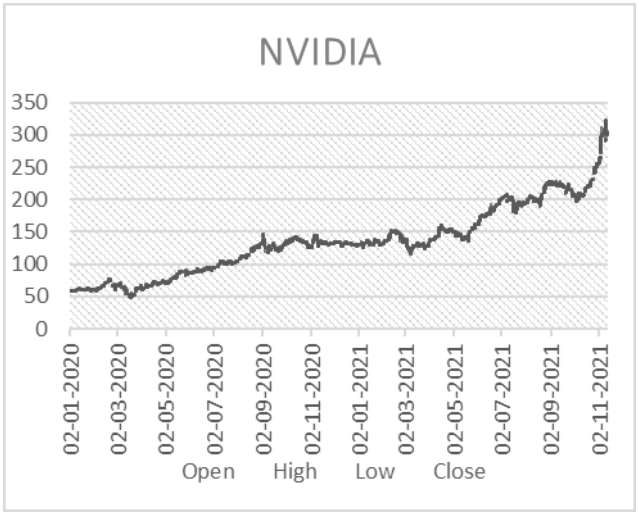

Figure 8 Nvidia Stock Price Analysis

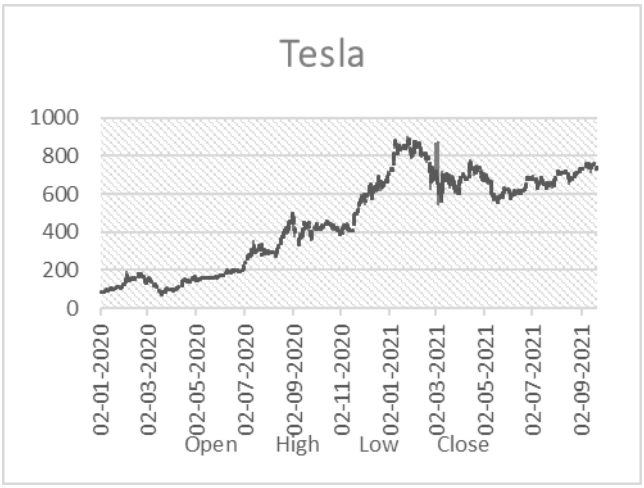

Figure 9 Tesla Stock Price Analysis

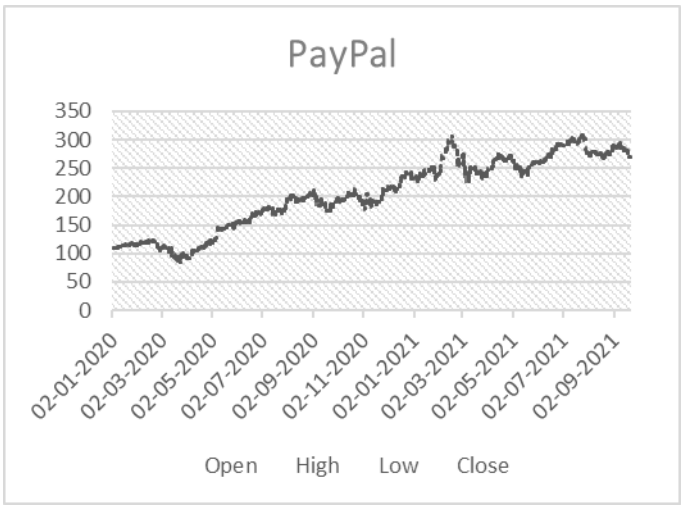

Figure 10 Tesla Stock Price Analysis

C. Data Agregation

We have merged the Financial Stock data of the five company's data and the tweeter sentimental data, further merging the data-wise total sentiment like (Data: 20-2-21 Positive: 5, Negative: 8, Neutral :3), which can be used easily to train the machine learning model. The final database has the Year, Month, Day, Stock Name, Positive, Negative, Neutral, Total Tweets, Volume, Open, High, and Low, further, we have label encoded the stock name like (Apple, Microsoft, Tesla, PayPal, Nvidia : 0, 1,2,3,4) and finally normalize the data by converting every attribute value between 0 and 1.

In our final Combination of Microeconomic and Financial Datasets of Five companies there are approximate 2983 rows and 13 columns.

IV. IMPLEMENTATION

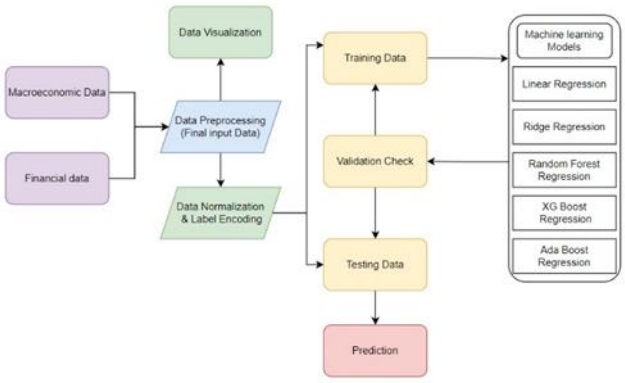

Figure 11 Flowchart

As the proposed work (Figure 13) compares various machine learning models, our first step will be to import the necessary libraries, i.e., pandas, numpy, sklearn, math, metrics, Standard Scalar, and many more. After importing important libraries the next step will be to read the dataset. The dataset is a combination of twitter data and financial data. The data pre-processing of twitter data and financial data is already discussed in the above section in detail. The dataset has financial data of 5 companies. The company names are labels in the dataset so it has to be converted to numeric data. For converting the labels into numeric data we used the sklearn label encoding. Data is then visualized by using the matplotlib for better visualization of data as shown in Figure 14, Figure 15, Figure 12.

Figure 12 Volume vs Close price Graph

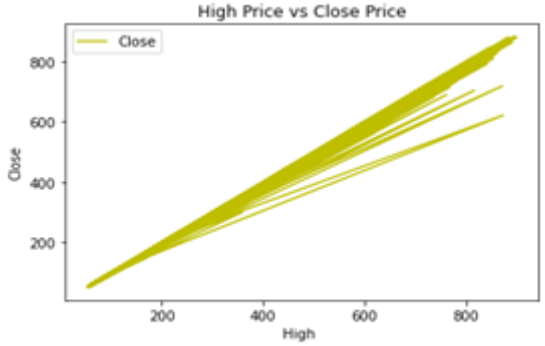

Figure 13 High vs close Price Graph

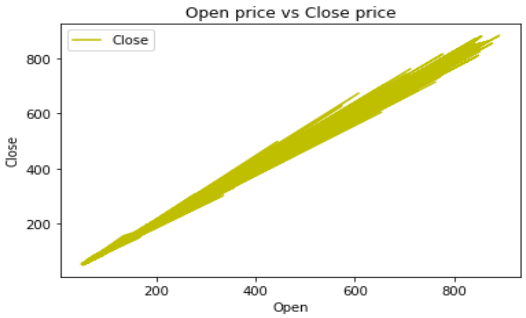

Figure 14 Open vs Close Price Graph

The next phase in the data pre-processing process is data normalisation. Data are transformed during normalisation to make them either dimensionless or have comparable distributions. Other names for this normalising procedure include standardisation, feature scaling, etc. Any machine learning and model fitting application must include normalisation as a crucial stage in the data pre-processing process.

The division of the data into training and testing data is a crucial step before fitting the data to the model. To partition the data more effectively, time series split is utilised. The standard scaler is then used to convert the X train data.Standard Scaler is an important technique that is mainly performed as a pre-processing step before many machine learning models to standardize the range of functionality of the input dataset. Standard Scaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

The data is finally integrated into several machine learning models. In this system, five different models—Linear regression, Ridge regression, Random Forest regression, XG-Boost, and AdaBoost Regression—are employed. After evaluating multiple models on this dataset, these five models were determined to be the best ones. To determine which performance metric best fits the model for the suggested system, it is put to the test. The key performance indicators examined on the model are R-squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).Root Mean Square Error (RMSE) is the residuals' standard deviation (prediction errors). The distance between the data points and the regression line is measured by the residuals, and the spread of these residuals is determined by the RMSE. In other words, it provides information on how tightly the data is clustered around the line of best fit. The size of the discrepancy between an observation's predicted value and its actual value is referred to as the absolute error. The amount of variation for the dependent variable that is explained by the independent variable or variables in the regression model is expressed statistically as R-squared (R2). These performance indicators are used to compare different models.

V. Machine Learning Models

A. *Linear Regression Model*

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one that you use to make a prediction about the value of another variable.

A linear regression line equation is written in the form of :-

$$Y = a + bX$$

Equation 1 Line Equation

In Equation 1 where X is the independent variable, displayed as a plot along the x-axis. The dependent variable, Y, is represented by a line on the y-axis. A is the intercept (the value of y when x = 0), and b is the line's slope. Here, we can find the intercept a and the slope (b) of the line plotted in a scatter plot. (**Error! Reference source not found.**) . We have also plotted a graph on basis of our datapoints in dataset as shown in **Error! Reference source not found.**.

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Equation 1 Values of a & b

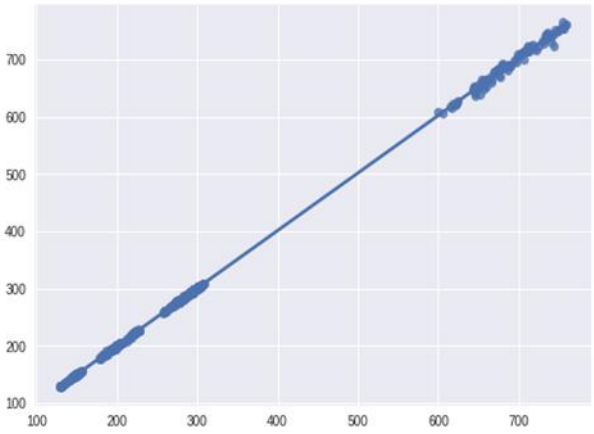

Figure 16 Linear Regression Model

B. Ridge Regression Model

Ridge regression is a model-tuning technique used to examine any multicollinear data. This technique carries out L2 regularisation. The projected values are far from the true values when the multicollinearity problem is present because the least squares are unbiased and the variances are high.

In ridge regression, the cost function also includes a factor known as the "sum of squares of the coefficients." In essence, ridge regression seeks to reduce both the total of the error term and the sum of the squares of the coefficients that we seek to determine. The "regularization term" is the sum of the squares of the coefficients, and the regularization coefficient is shown by, as illustrated. in Equation 5.

$$\underset{\beta_0,\beta}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Equation 2 Ridge Cost Functions

The coefficient values of the above cost function are determined by the following closed form solution (Equation 6). We have also plotted graph on basis of our datapoints in dataset by using Ridge regression model as shown in Figure 17.

$$\beta = (X^T X + \lambda I)^{-1} X^T Y$$

Equation 3 Ridge equation

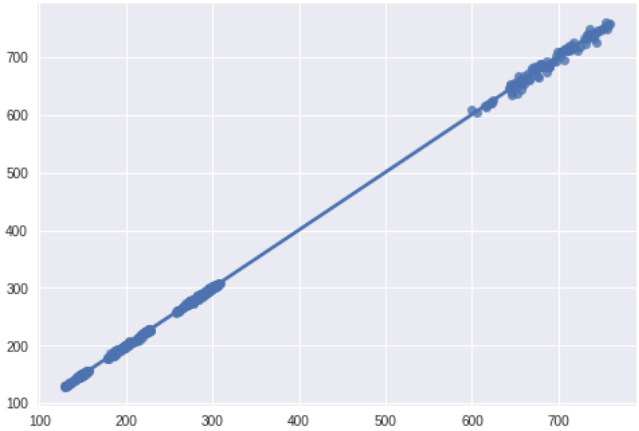

Figure 17 Ridge Regression Model

C. *Random Forest Regression Model*

Using several decision trees and a method called Bootstrap and Aggregation, sometimes known as bagging, Random Forest is an ensemble methodology capable of handling both regression and classification problems. Rather than depending just on one decision tree to determine the final result, the primary idea is to merge numerous decision trees.The importance for each feature on a decision tree is then calculated as shown in Equation 3. fi sub(i)= the importance of feature i[ni]. sub(j)= the importance of node j.

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\in all\ nodes} ni_k}$$

Equation 4 importance for each feature

$$normfi_i = \frac{fi_i}{\sum_{j\in all\ features} fi_j}$$

Equation 5 Finding Norms

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values as shown in Equation 4. At the Random Forest level, the ultimate feature relevance is determined by its average over all trees. The total number of trees is divided by the sum of the feature's importance rating on each tree as shown in Equation 7 . Graph for Random Forest Regression is shown in Figure 18.

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

Equation 6 Random Forest Regression

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model.
- normfi sub(ij)= the normalized feature importance for i in tree j.
- T = total number of trees

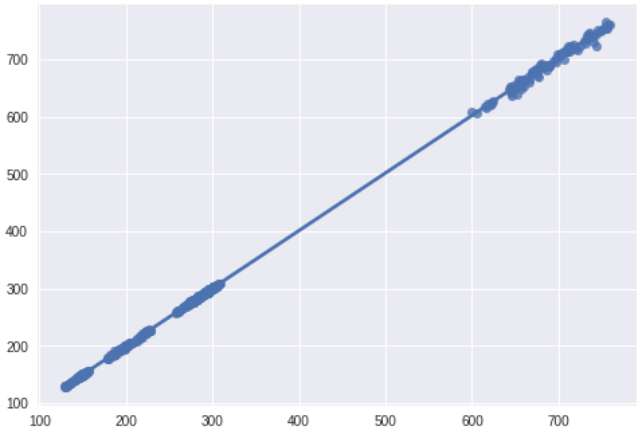

Figure 18 Random Forest Regression

D. XG Boost Regression Model

A gradient boosting framework is used by the decision tree-based ensemble machine learning technique known as XG Boost. Artificial neural networks frequently outperform all other algorithms or frameworks in prediction issues involving unstructured data (pictures, text, etc.). However, decision tree-based algorithms are now regarded as best in class when applied to small to medium structured/tabular data. We use I to represent each example in a dataset if it contains "n" examples, which corresponds to n rows. By minimising the following values, XG Boost builds trees using the loss function. in Equation 8.

We have also plotted graph as shown in Figure 19.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

Equation 7 XG Boost Equation

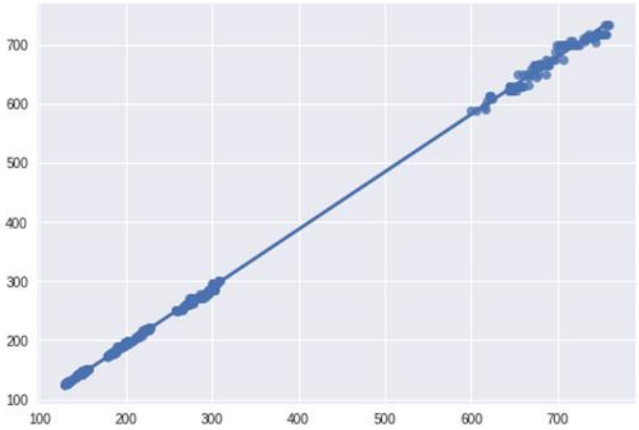

Figure 19 XG Boost Regression Model

E. Ada Boost Regression Model

One of the first boosting algorithms to be applied in solution processes was Ada Boost. Multiple "weak classifiers" can be combined into a single "strong classifier" with the aid of Ada Boost. Decision trees with a single split, or "decision stubs," are the weaker learners in AdaBoost. AdaBoost works by giving harder-to-classify instances more weight and undervaluing examples that have already undergone thorough processing. The AdaBoost algorithms can be applied to problems involving classification and regression..

Suppose we are given training data {(xi , yi)} N i = 1, where xi ∈ R K and yi ∈ {−1, 1}. And suppose we are given a (potentially large) number of weak classifiers, denoted fm(x) ∈ {−1, 1}, and a 0-1 loss function I, defined as shown in Equation 8. The final classifier is built using a linear combination of the weak classifiers after learning. as shown in Equation 9. AdaBoost is essentially a greedy algorithm that adds and optimises the weights for one weak classifier at a time in order to construct a "strong classifier," or g(x), progressively. Additionally, we created the graph you see in Figure 20.

$$I(f_m(\mathbf{x}), y) = \begin{cases} 0 & \text{if} f_m(\mathbf{x}_i) = y_i \\ 1 & \text{if} f_m(\mathbf{x}_i) \neq y_i \end{cases}$$

Equation 8 Loss Function

$$g(\mathbf{x}) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})\right)$$

Equation 9 Ada Boost Regression

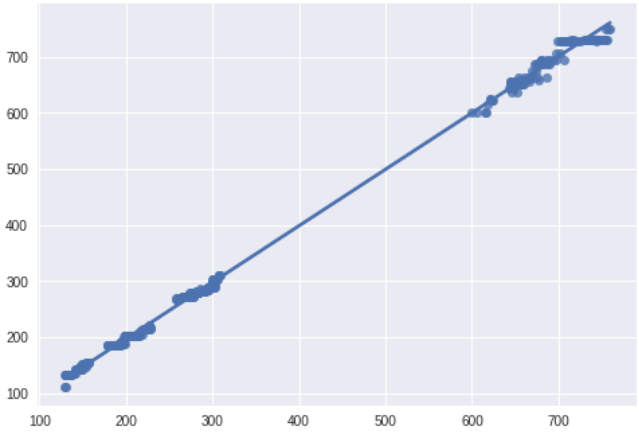

Figure 20 Ada boost Regression Model

VI. Results ans Analysis

The given table Table 1 below is the results of the Linear regression, Ridge Regression, Random Forest Regression, XG Boost and AdaBoost models . The Predicted close price and actual close price of that data entry is displayed and compared.

Table 1 Results

Model Name	Root Mean Square Error	Mean Avg. Error	R-Squared
Ridge Regression	3.120211	1.397399	0.999868
Linear Regression	3.130533	1.387486	0.999868
Random Forest Regression	3.130533	1.387486	0.999868

XG Boost Regression	3.909600	5.735600	0.999100
AdaBoost Regression	7.177600	5.291600	0.998600

The table given below (Table 2) shows the comparison of all models based on the Root Mean Squared Error, Mean Average Error and R-squared value. Clearly it is visible the values are very close to each other indicating the performance of the models.

Table 2 Actual vs Predicted Prices

Actual Close Prize	Ridge	Linear	Random Forest	XG Boost	Ada Boost
	Predicted Price				
178.897	177.620	178.012	178.012	171.126	185.620
273.190	272.404	273.326	273.326	259.086	272.056
129.436	127.582	127.702	127.702	124.211	110.721
259.223	258.352	258.832	258.832	249.013	268.181
615.090	615.996	618.376	618.376	594.965	600.942

The Comparison of the models is plotted as a bar graph for better understanding of the data as shown in Figure 21.

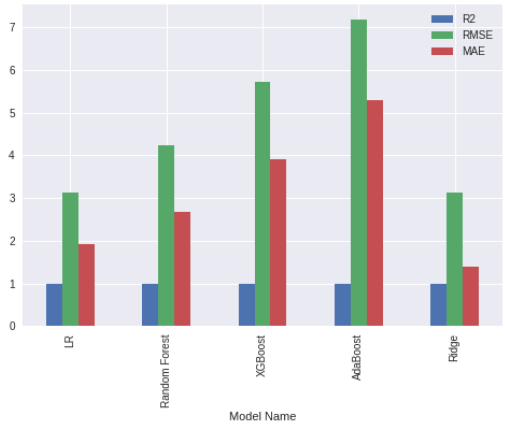

Figure 21 Comparison Graph

Finally, the Ridge regression came out to be the best model among all the other models. Its RMSE score is 3.120211, and R-squared value is 0.999868 which is better than other models by very less margin.

VII. CONCLUSION

In this study, five machine learning models are tested on a unique dataset, which is a combination of twitter dataset and financial dataset. Before training the models, all pre-processing steps are performed for getting better results. The given methodology is a healthy comparison between various machine learning models based on Root Mean Squared Error, Mean Squared Error and R-Square values for better comparison. Linear Regression, Ridge Regression, Random Forest Regression, XG Boost and Ada Boost models are used to test the data. Ridge Regression model gives the best accuracy with the RMSE Score of 3.120211 and R-Squared Value of 0.999868 and proves to be the best fit by the study followed by Linear Regression with the RMSE Score of 3.130533 and R-Squared Value of 0.999868. With more diverse data from various sources related to our study will definitely help the model to predict values and get better results. The study focuses on predicting the closing price of stocks of five major companies by analysing the twitter news trends related to macroeconomics. But if we compare all the Models, the performance metrics are all very close to each other and thus the output graphs of all the models are very similar to each other as the difference between them is very less.

VIII. FUTURE SCOPE

The current study focuses on predicting the closing price of stocks and gives out very accurate results. One of the scope of this study will be to predict the stock prices while simultaneously looking into the current trending macroeconomic news and the social media news including twitter , Instagram and other important social media apps. The study at this instance focusses on predicting the closing price for five major companies, so , there is a scope of covering large number of stocks and predict the same for them.

REFERENCES

[1] Jasic, T., & Wood, D. (2004). The profitability of daily stock market indices trades based on neural network predictions: Case study for the S&P 500, the DAX, the TOPIX and the FTSE in the period 1965–1999. Applied Financial Economics, 14(4), 285-297.

[2] Chavan, P. S., & Patil, S. T. (2013). Parameters for stock market prediction. International Journal of Computer Technology and Applications, 4(2), 337.

[3] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259-268.

[4] Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.

[5] Das, Shom Prasad, and Sudarsan Padhy. "Support vector machines for prediction of futures prices in Indian stock market." International Journal of Computer Applications 41, no. 3 (2012).

[6] Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Standford University, CS229 (2011 http://cs229. stanford.edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSentimentAnalysis. pdf) 15 (2012): 2352.

[7] Vijh, Mehar, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. "Stock closing price prediction using machine learning techniques." Procedia computer science 167 (2020): 599-606.

[8] Nti, Kofi Opoku and Adekoya, Adebayo and Weyori, Benjamin, Random Forest Based Feature Selection of Macroeconomic Variables for Stock Market Prediction (July 19, 2019). American Journal of Applied Sciences 2019, 16 (7): 200.212 DOI: 10.3844/ajassp.2019.200.212 .

[9] N. Waduge and U. Ganegoda, "Forecasting Stock Price of a Company Considering Macroeconomic Effect from News Events," 2018 3rd International Conference on Information Technology Research (ICITR), 2018, pp. 1-5, doi: 10.1109/ICITR.2018.8736133.

[10] Duan, Yuejiao, John W. Goodell, Haoran Li, and Xinming Li. "Assessing machine learning for forecasting economic risk: Evidence from an expanded Chinese financial information set." Finance Research Letters 46 (2022): 102273.

[11] Weng, Bin, Waldyn Martinez, Yao-Te Tsai, Chen Li, Lin Lu, James R. Barth, and Fadel M. Megahed. "Macroeconomic indicators alone can predict the monthly closing price of major US indices: Insights from artificial intelligence, time-series analysis and hybrid models." Applied Soft Computing 71 (2018): 685-697.

[12] J. Bollen and H. Mao. Twitter mood as a stock market predictor. Computer, 44(10):91–94, 2011.